Analysis Of Factors Impacting AI & Data Science Project

NLP Final Project

Prepared for Natural Language Processing Class

12/08/2022

By: Naibo(Ray) Hu

Agenda

01 Executive Summary

02 Data Profile & Preprocessing

03 Detect Major Topics

04 Sentiment Analysis & Timeline

05 Analysis of failing AI projects

06 Analysis of successful AI projects

07 Targeted Sentiment Analysis

08 Recommendation

Executive Summary

Business Problem

Nowadays, **automation is widely used** in various industries. Companies utilize AI to reduce costs, time, and waste as well as increase productivity, reduce mistakes, and control all the processes of the business in real time.

However, there are many **risks and disadvantages associated with AI** that can negatively impact human.

Goal of Analysis

Identify the underlying **reasons for successes and failures** in data science initiatives by extracting meaningful insights from unstructured text

Provide actionable recommendations on what can be done to increase the success rates of the data science capabilities

Insights & Recommendation

Key reasons for **failing** data science initiatives involve ***ethical issues, employment insecurity, crime activity, stock price decline, lack of funding, and project shutdown.***

To increase **success rate** of data science projects:

- Companies should ***invest in R&D for new product innovation, develop cutting edge technologies, improve cybersecurity, and prioritize projects properly.***
- Government should ***allocate more funds to support tech companies***

Data Profile & Data Preprocessing

Data profile

Name	AI & Data Science Project Data
Dimensions	5 variables 199,538 rows
Variables name	Url, date, language, title, text

After data preprocessing, **168,797 rows** are left.

The shortest article has 5 words, and the longest article has 21,119 words.

Distribution of article length

	length
count	168797.000000
mean	925.124439
std	904.795189
min	5.000000
25%	475.000000
50%	749.000000
75%	1146.000000
max	21119.000000

Cleaning steps:

Clean up noise

Remove new line, url, mention, tag, consecutive whitespace in article text and title

Remove remnants of web claws

- '\xa0SAP', '\xa0National', '\xa0'
- , '\t'
- ,'\r'
- '\|'

Discard irrelevant articles

Filter out articles contain words:

'AI', 'ai', 'artificial intelligence', 'Artificial Intelligence', 'ARTIFICIAL INTELLIGENCE', 'Data Science', 'data science', 'DATA SCIENCE'

Drop duplicates

Drop duplicated rows based on article text and titles

Run BERTopic Modeling on all articles and detect major topics

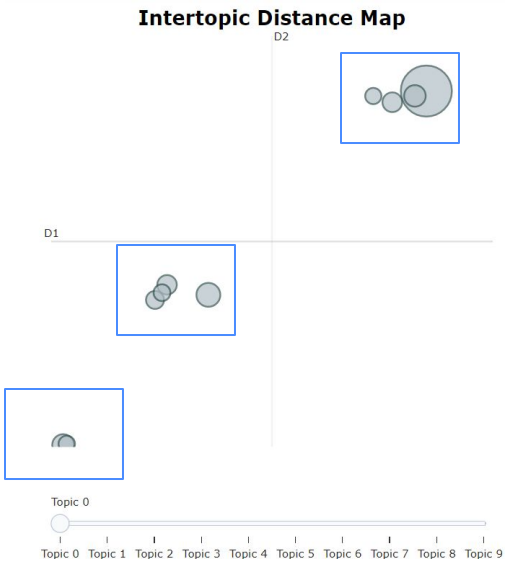
Since data is related to AI, **drop following keywords** to obtain better results

- "'ai', 'artificial intelligence', 'data', 'data science', 'science', 'intelligence', 'artificial', 'intelligence'

Fit BERTopic model with **number of topics = 10**

- 10 topics can be clustered to 3 main groups

Topic -1 refers to outliers, which are dropped



Topic overview

Topic	Count	Name
-1	128723	-1_new_news_technology_us
0	17347	0_market_report_analysis_growth
1	3796	1_new_star_reveals_first
2	3104	2_products_public_releasesign_releasesign uplog
3	3102	3_platform_technology_credit_digital
4	2576	4_news_weather_valley_schedulewdpntv
5	2527	5_shares_traded_inc_matrix
6	2152	6_nvidia_zdnet_gpu_new
7	1897	7_mar mar_mar_jan jan_mar mar mar
8	1812	8_platform_solutions_technology_companies
9	1761	9_products_news_public_releasesign

Topic 0 to Topic 9 represent major topics of AI articles. Topic 0 (market report & growth analysis) is the most frequently mentioned topic

Further analyze the 10 topics generated from BERTopic model

To further understand articles' major topics, I look into top keywords related to each topic and summarize text by topic.

Top keywords by topic

	0	1	2	3	4	5	6	7	8	9
Topic_0	market	report	analysis	growth	global	forecast	research	industry	key	players
Topic_1	new	star	reveals	first	show	says	dress	shows	one	looks
Topic_2	products	public	releasesign	releasesign uplog	uplog	services	news	business	industry	technology
Topic_3	platform	technology	credit	digital	software	solutions	new	zest	content	company
Topic_4	news	weather	valley	schedulewdpntv	us	scores	dashboard	lehigh	business	solutions
Topic_5	shares	traded	inc	matrix	dollar	price	etf	stock	dollar trades	trades
Topic_6	nvidia	zdnet	gpu	new	edge	computing	cloud	performance	jetson	gpus
Topic_7	mar mar	mar	jan jan	mar mar mar	jan	jan jan jan	feb feb	may may	jun jun	feb
Topic_8	platform	solutions	technology	companies	business	insights	company	industrial	new	canvass
Topic_9	products	news	public	releasesign	releasesign uplog	uplog	services	consumer	business	industry

Major topics associated with AI projects include:

- Market analysis & forecasting in various industries
- Technology product, service and company
- Algorithms (cloud computing, ML)
- Finance & stock price
- Customer review & insights

Topic	Text Summarization
0	Market analysis & growth forecast
1	Startups
2	Technology product & service
3	Technology, software, digital company
4	Weather prediction
5	Finance & stock
6	Cloud computing & GPU
7	Date
8	Technology insights & news
9	Consumer review

Run LDA Topic Modeling on all articles and detect major topics

Coherence score by topic

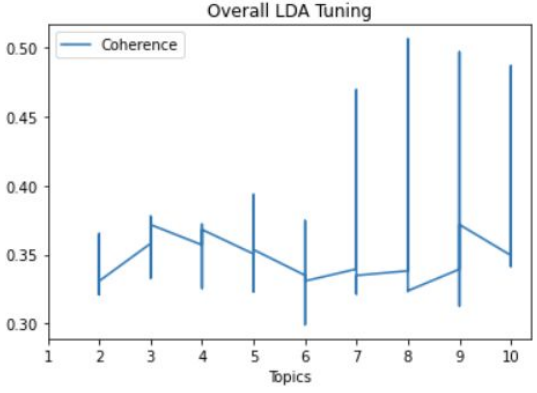
Topics	Alpha	Beta	Coherence
8	0.01	0.91	0.506472
8	0.91	0.91	0.504507
9	0.61	0.91	0.497020
10	symmetric	0.91	0.486851
10	0.01	0.91	0.484490
7	asymmetric	0.91	0.469683
10	asymmetric	0.91	0.467886
9	0.91	0.91	0.462129
10	0.91	0.91	0.458198
8	0.61	0.91	0.457577

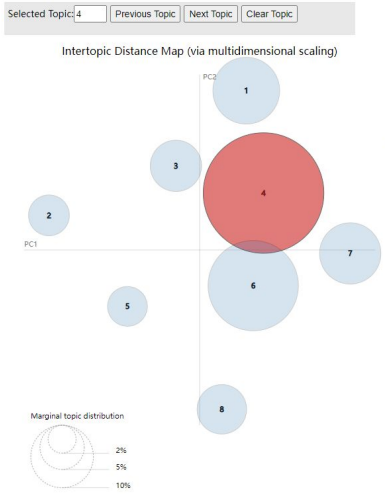
Besides BERTopic, I also run LDA model to detect major topics. The best model contains **8 topics** with the coherence score of 0.506.

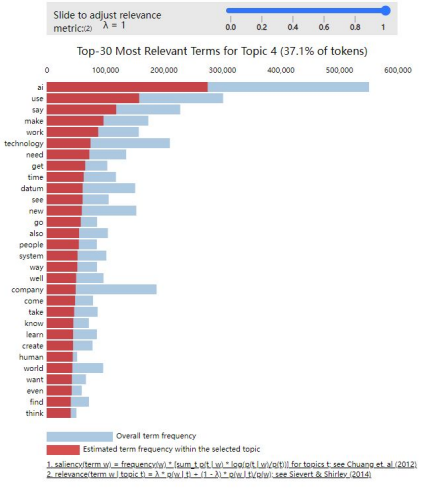
LDA Model Topic

(0, '0.007*"say" + 0.005*"show" + 0.004*"new" + 0.004*"make" + 0.003*"reveal" + 0.003*"use" + 0.003*"ai" + 0.003*"take" + 0.003*"day" + 0.003*"year"')
(1, '0.013*"trade" + 0.009*"dollar" + 0.008*"ai" + 0.005*"use" + 0.005*"company" + 0.004*"share" + 0.003*"technology" + 0.003*"stock" + 0.002*"exchange" + 0.002*"high"')
(2, '0.009*"say" + 0.006*"company" + 0.005*"use" + 0.005*"ai" + 0.003*"technology" + 0.003*"new" + 0.003*"rights_reserve" + 0.002*"year" + 0.002*"make" + 0.002*"work"')
(3, '0.014*"ai" + 0.008*"use" + 0.006*"say" + 0.005*"make" + 0.005*"work" + 0.004*"technology" + 0.004*"need" + 0.003*"get" + 0.003*"time" + 0.003*"datum"')
(4, '0.009*"ai" + 0.003*"say" + 0.003*"technology" + 0.003*"use" + 0.002*"make" + 0.002*"company" + 0.002*"work" + 0.002*"datum" + 0.002*"world" + 0.002*"power"')
(5, '0.012*"ai" + 0.007*"technology" + 0.007*"company" + 0.005*"use" + 0.005*"business" + 0.005*"datum" + 0.004*"customer" + 0.004*"solution" + 0.004*"service" + 0.003*"platform"')
(6, '0.008*"ai" + 0.005*"use" + 0.005*"patient" + 0.003*"technology" + 0.003*"news" + 0.002*"datum" + 0.002*"company" + 0.002*"overview" + 0.002*"solution" + 0.002*"work"')
(7, '0.021*"market" + 0.009*"ai" + 0.009*"report" + 0.006*"analysis" + 0.006*"industry" + 0.005*"forecast" + 0.005*"growth" + 0.004*"global" + 0.004*"technology" + 0.003*"market_report"')

LDA model output aligns with BERTopic results







Accordion to LDA, major topics include

- Stock price
- Technology product & company
- Industry-related forecast & analysis

Train Custom classifier for sentiment analysis

Use Yelp data to train classifier

Yelp data example

- label 1: positive sentiment,
- label 0: negative sentiment

	text	label	lang
	I love Deagan's. I do. I really do. The atmosp...	1	en
	I love the classes at this gym. Zumba and. Rad...	1	en

Yelp data has a total of **255,717 rows**.

Steps:

1. Train-test data split (85% train 25% test)
2. Create ML pipelines
   a. Apply TfidfVectorizer to vectorize and normalize data
   b. Remove stopwords
   c. Make ngram (1 - 3)
   d. Train support vector machine & logistic regression classifiers

Apply **logistic regression** classifier to AI data and predict article sentiment

Logistic regression has a higher accuracy and f1 score than SVM.

Support Vector Machine (SVM) model performance

	precision	recall	f1-score	support
0	0.95	0.97	0.96	19199
1	0.97	0.95	0.96	19159
accuracy			0.96	38358
macro avg	0.96	0.96	0.96	38358
weighted avg	0.96	0.96	0.96	38358

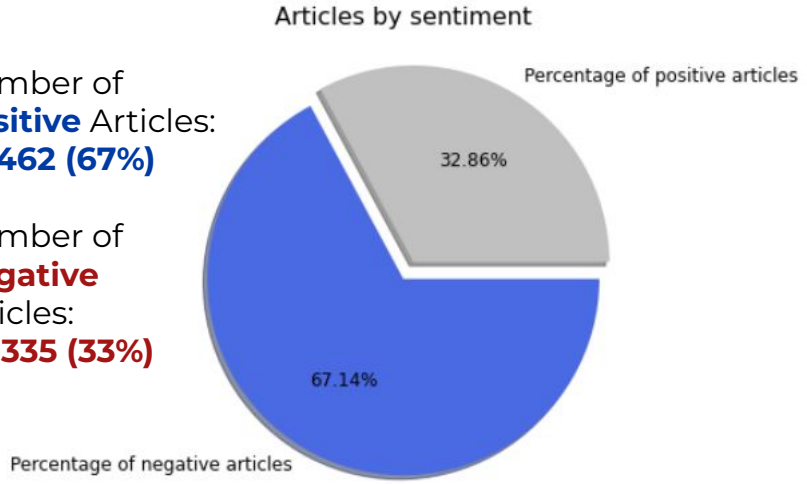
Logistic regression model performance

	precision	recall	f1-score	support
0	0.96	0.97	0.97	19199
1	0.97	0.96	0.97	19159
accuracy			0.97	38358
macro avg	0.97	0.97	0.97	38358
weighted avg	0.97	0.97	0.97	38358

Majority articles have negative sentiment

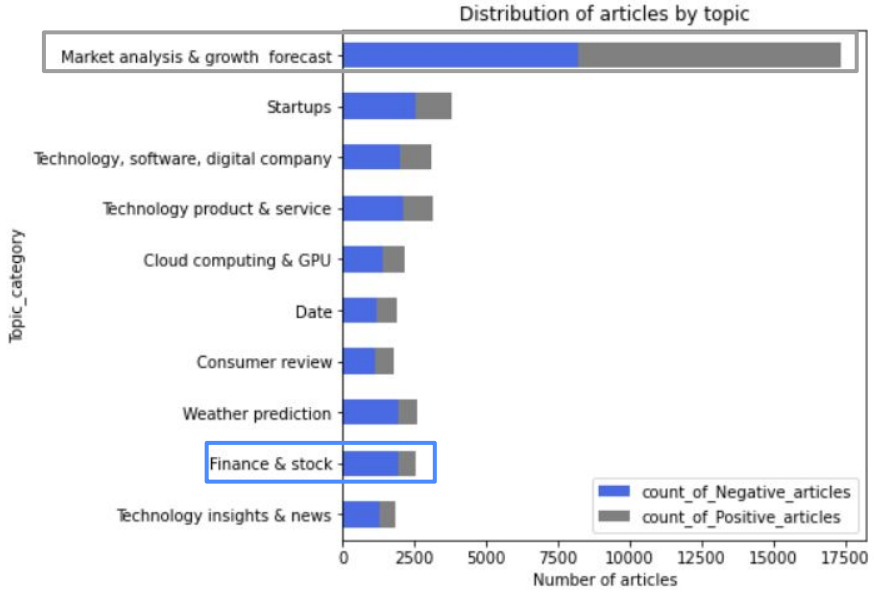
Number of **Positive** Articles: **55,462 (67%)**

Number of **Negative** Articles: **113,335 (33%)**



AI Data with sentiment

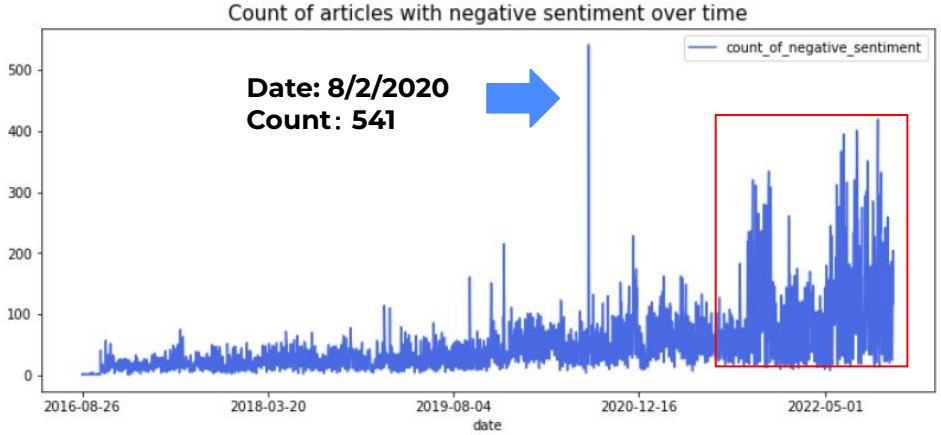
date	clean_title	clean_text	sentiment	sentiment_category
2022-03-09	Gender Bias in A...	Gender Bias in A...	0	Negative
2017-07-04	Big pharma turns...	Big pharma turns...	0	Negative
2018-03-13	Amazon HQ2 Winne...	Amazon HQ2 Winne...	0	Negative
2019-02-13	Trump's 'America...	Trump's 'America...	0	Negative
2018-03-27	Xiaomi to relea...	Xiaomi to releas...	0	Negative



According to the stack bar chart, the most positive topic is **market analysis & growth forecast**. The most negative topic is **Finance & Stock**.

Besides the topic of market analysis & growth forecast, the rest topics have more negative sentiment articles.

Negative Sentiment Analysis Over Time

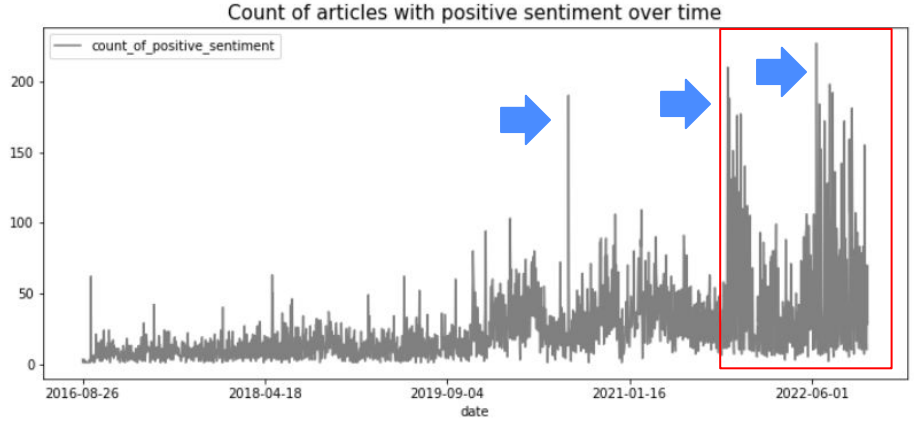


Articles with Negative Sentiment

- Number of negative articles peaks on Aug 2, 2020.
- Number of negative articles increase exponentially after 2022.

By plotting the word cloud based on articles from 8/2/2020, I find that majority negative articles are associated with **weather forecasting, COVID, cloud service, App, and 5G network**.

Positive Sentiment Analysis Over Time



Articles with Positive Sentiment

- Number of positive articles peaks in late 2021, mid 2022, and mid 2020.
- Number of positive articles increase exponentially after 2022.

By plotting the word cloud based on articles from 6/14/2022, 10/14/2021, 8/2/2020, I find that majority positive articles are associated with **DALL (deep learning algorithm), Amazon E2 computing service, Image processing, and supply chain.**

Analysis of article text with Negative sentiment

Entity Identification

Spacy organization

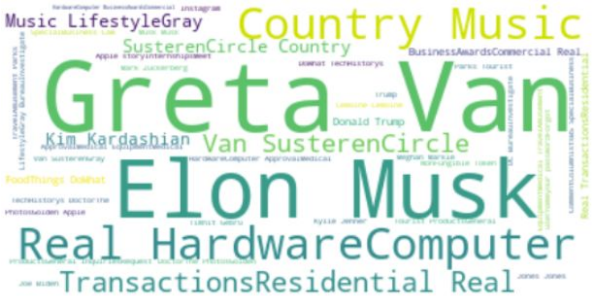
Entities	Count
Google	42475
Facebook	19445
Microsoft	17626
Amazon	15257
IBM	12235
Apple	8941
Tesla	6951
Intel	6863
NewscastsPress	6847
Googles	6753
EU	6020
ReleaseSign UpLog	5446
Nvidia	5299
TechConsumer	4914
OfficesSend	4833
RD	4802
BrandVoice	4663
Samsung	4598
CasinosHotels	4587
Huawei	4446

Spacy Person

Entities	Count
HardwareComputer	7990
Trump	5049
Elon Musk	4653
TransactionsResidential Real	4346
UsContact	3927
Greta Van SusterenCircle Country Music LifestyleGray	3805
Forbes	3535
Musk	3418
storyInternshipsMeet	3108
ReleaseSign	3043
BusinessAwardsCommercial Real	2930
DemoEditorial	2490
Kim Kardashian	2450
Jones	2447
Biden	2445
ApprovalMedical EquipmentMedical	2428
BureauInvestigate	2331
DoctorThe	2304
TravelAmusement Parks Tourist	2193
ProductsGeneral InquiriesRequest	2192

Word Cloud (Organization)

Word Cloud (Person)



The following entities are associated with **negative** sentiment

Company:

- Google, Microsoft, IBM, Facebook, Amazon, etc

People:

- Elon Musk, Greta Van, Kim Kardashian, Trump, and Biden

Technology & Product:

- Hardware computer, Country Music

Analysis of article title with Negative sentiment

Entity Identification

Spacy organization

Entities	Count
Google	5029
Microsoft	2741
IBM	2569
Facebook	1510
Amazon	1282
Intel	1159
TechRadar	1054
Nvidia	650
IBM Corporation	586
Courier	565
Microsoft Corporation	563
Reuters	481
SAP	462
ZDNet	450
Apple	438
Tesla	431
NVIDIA	418
The Bisouv Network	410
Ford	397
Intel Corporation	394

Spacy Person

Entities	Count
Elon Musk	454
Murphy	267
Galus Australis	154
Timnit Gebru	131
Biden	91
Forbes	80
Monroe Scoop	79
Trump	79
Zuckerberg	79
| Engadget	78
Benzinga	78
APAC	76
CauseACTION Clarion	76
Siri	73
DataRobot	72
FN Meka	72
Mark Zuckerberg	65
Musk	65
PreciTaste	64
Standigm Signs MOU	59

Word Cloud (Organization)

Word Cloud (Person)

The following entities are associated with **negative** sentiment

Company:

- Google, Microsoft, IBM, Facebook, Amazon Web, Forbes, etc

People:

- Elon Musk, Timnit Gebru, Mark Zuckerberg, Trump, Biden

Identify top reasons for failing data science initiatives

Step 1: From articles with Negative sentiment, filter contents based on top organizations and people

- Google, Microsoft, IBM, Facebook, Amazon, Forbes, Elon Musk, Timnit Gebru, Zuckerberg, Greta Van, Trump, and Biden

Step 2: Apply BERTopic model to identify major topics and reasons

Reason(Topic)	Article text example	Corrective Action
Stock price decline	• Tesla Stock News and Forecast: TSLA stock loses direction as AI head departs • Samsung vows 5G and AI lead, shareholders lambast low stock price	Improve company **brand images and reputation** to make people have more confidence in company stock
Cloud computing shut-down	• Huawei is closing its cloud and AI business group after only a year • Facebook Shuts Down AI Experiment • EU Data Watchdogs Call for Ban on Facial Recognition Through AI	**Prioritize projects**. Manage data **on-premise** instead of cloud, as cloud operation can be expensive
Crime	• Dentist charged by SEC for digital token project fraud • Thieves are now using AI deepfakes to trick companies into sending them money	**Improve cyber security** and prevent criminal actions
Ethical issues	• Google AI researcher's exit sparks ethics, bias concerns • The age of AI in healthcare: disrupting efficiency & impacting ethics • Elon Musk Warns 'Greatest Risk' To Civilization Is Artificial Intelligence	Continue **human review of AI-assisted decision-making**. Implement informed consent when necessary.
Employment insecurity	• Google fires AI manager who protested her peer's departure • Google fires engineer who contended its AI technology was sentient • Nearly 9 million British jobs could be lost to AI by 2030	**Improve employee benefits** and optimize hiring process
Tech startups lack funding	• Council Post: Seeking Investors For Your AI Startup? • Trump's 'American Artificial Intelligence Initiative' Needs Money	Seek funding from **venture capitals**. Apply **bank loans**.

Analysis of article text with positive sentiment

Entity Identification

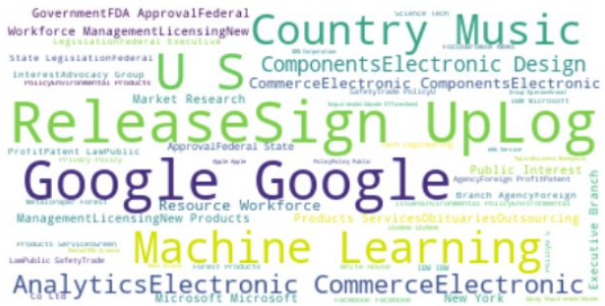
Spacy organization

Entities	Count
Google	15559
Microsoft	8517
IBM	8349
Facebook	7863
Apple	5707
Amazon	5523
Intel	4056
Samsung	3565
Tesla	3287
Googles	2851
ReleaseSign UpLog	2836
NewscastsPress	2805
Nvidia	2656
TechConsumer	2514
OfficesSend	2489
CasinosHotels	2403
EU	2378
USNew	2341
Huawei	2322

Spacy Person

Entities	Count
HardwareComputer	4192
Trump	2344
TransactionsResidential Real	2260
Elon Musk	1769
ReleaseSign	1647
UsContact	1618
Greta Van SusterenCircle Country Music LifestyleGray	1588
BusinessAwardsCommercial Real	1534
Kim Kardashian	1414
Instagram	1317
DemoEditorial	1291
ApprovalMedical EquipmentMedical	1262
Musk	1216
TravelAmusement Parks Tourist	1129
ProductsGeneral InquiriesRequest	1097
Forbes	976
BureauInvestigate	909
Biden	867
Siri	859
Donald Trump	835

Spacy organization Word Cloud



Spacy person Word Cloud

The following entities are frequently associated with **positive** sentiment

Company:

- Google, Microsoft, IBM, Facebook, Amazon, etc

People:

- Elon Musk, Greta Van, Kim Kardashian, Trump, and Biden

Technology & Product:

- Hardware computer, Machine learning, Transactions residential, Instagram

Analysis of article title with positive sentiment

Entity Identification

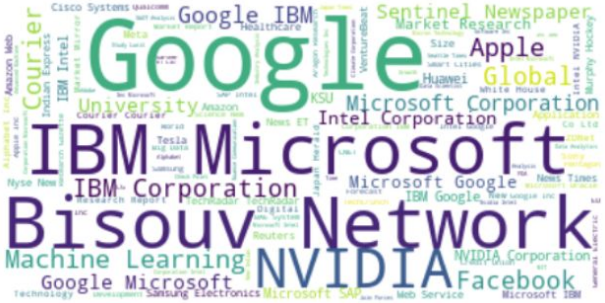
Spacy organization

Entities	Count
Google	2502
IBM	2376
Microsoft	2169
Intel	1197
Courier	674
The Bisouv Network	597
TechRadar	569
Apple	554
Samsung	465
SAP	451
Nvidia	442
Facebook	431
NVIDIA	413
IBM Corporation	331
Oracle	291
Amazon	289
Sony	285
Huawei	275
KSU	272
The Sentinel Newspaper	271

Spacy Person

Entities	Count
Murphy	201
Monroe Scoop	149
Elon Musk	131
CE Mark	98
Galus Australis	98
Mark Zuckerberg	87
Rembrandt	82
Beethoven	64
John Deere	57
QYSEA	55
Leonardo Da Vinci	55
Albert Technologies	54
Tao	53
Janssen	53
Pope	51
Kelly Merton	48
Artemio Rimando	48
| Engadget	45
LandingLens	44
Zenarate	43

Spacy organization Word Cloud



Spacy person Word Cloud

The following entities are frequently associated with **positive** sentiment

Company:

- Google, Microsoft, IBM, Facebook, Amazon, Samsung, Nvidia, TechRadar, etc

People:

- Elon Musk, Mark Zuckerberg, Murphy, Beethoven, John Deere, Leonardo Da Vinci

What companies and government can do to improve outcome of data science initiative?

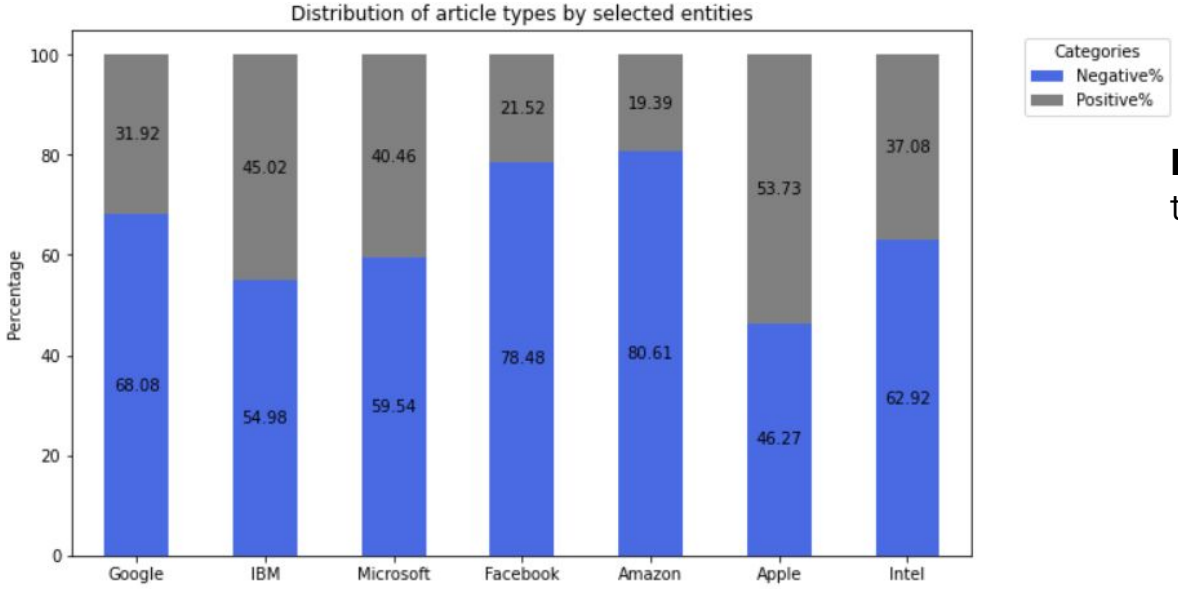
Step 1: From articles with Positive sentiment, filter contents based on top organizations and people

- Google, Microsoft, IBM, Facebook, Amazon, Samsung, Elon Musk, Mark Zuckerberg, Greta Van, Trump, and Biden

Step 2: Apply BERTopic model to identify major topics and reasons

Reason(Topic)	Article text example	Recommended Actions
Market analysis & Prediction	• Samsung: Digital health and AI will be driving the future growth of the Medical Device Industry: • Mobile Artificial Intelligence Future Demand Analysis, Industry Share, Top Key Vendors and Market Forecast upto 2018 – 2026	**Expand** high quality AI projects to various industry
Technology product innovation	• Samsung eyes screening of 150,000 Indians with unique AI camera • Using AI to find new pharmaceutical applications for natural products • Google Pixel 6 advert teases smart camera and AI features • Huawei targets Nvidia, Intel, Qualcomm with new AI chips • IBM unveils new chip designed to detect fraud with AI	**Invest money in R&D** and continuously emphasize on new product and service innovation.
Machine Learning	• Machine Learning Market is thriving worldwide by 2027 • The UK Medicines and Healthcare products Regulatory Agency (MHRA) selects Commonwealth Informatics Inc to explore the use of AI and Machine Learning across Safety Surveillance • Strategic Analysis to Understand the Competitive Outlook of Machine Learning in Retail Market	**Develop effective machine learning and deep learning algorithm** to improve project quality
Government support	• Biden to Tap Artificial Intelligence Expert as Top Business Diplomat • Trump's AI Initiative - Everything But Money	Government can **allocate more funds** to support tech companies innovation.

Targeted Sentiment: Analyze sentiment on selected organization entities



Facebook and Amazon receive the most **negative** sentiment.

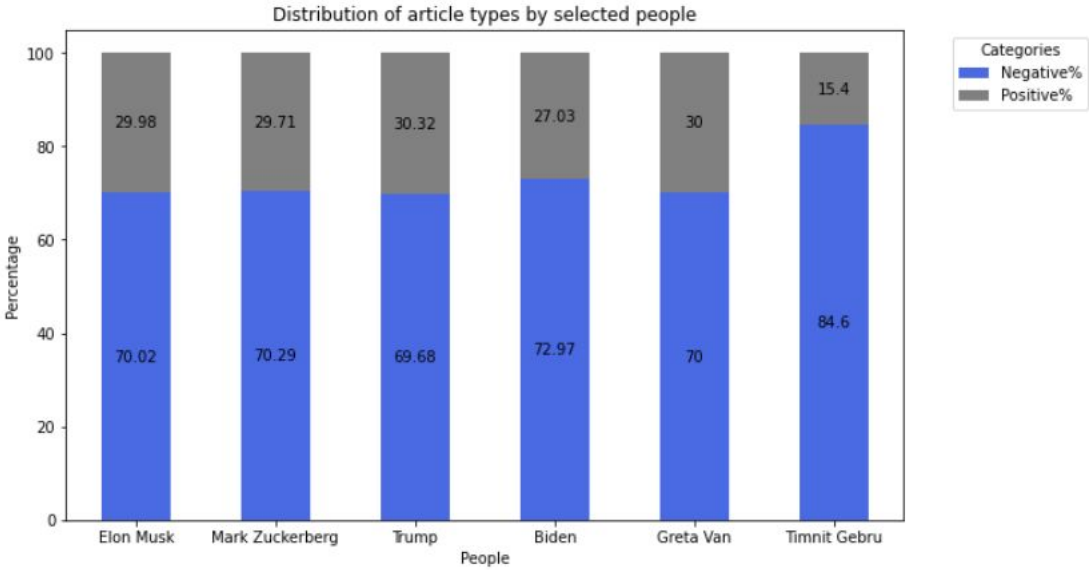
- **80%** of articles related to Amazon are negative
- **78%** of articles related to Facebook are negative

Apple has the most **positive** sentiment. **53%** of articles related to Apple are positive.

Steps:

1. Filter contents related to selected entities
2. Compute percentage of positive and negative articles for each entity

Targeted Sentiment: Analyze sentiment on selected people



Select frequently appeared people from articles

Timnit Gebru receives the most **negative** sentiment.

- **84.6%** of articles related to Timnit Gebru are negative
- Gebru has been recognized widely for her expertise in technology and artificial intelligence. She was **fired by Google after publishing a paper on the dangers of large language models**, like the ones that power the Google's search engine

People from **large tech companies** (Musk, Zuckerberg) and from **government** (Trump, Biden) also have more **negative** sentiment.

Recommendation

What types of mistakes business can avoid in data science space?

- Ethical issues
- Crime activity
- Employment insecurity
- Lack of funding
- Failure to prioritize project

Why businesses should invest in data science initiatives?

- Improve market research analysis & market prediction
- Stimulate product innovation
- Innovate cutting edge technology

Recommended Actions to increase success rate of AI project

Companies should

- Invest in R&D for new product innovation
- Develop cutting edge technologies
- Improve cybersecurity
- Prioritize projects properly
- Seek sufficient funding support
- Improve employee benefits

Government should

- Allocate more funds to support tech companies

Thanks